## e. Hypothesis Formulation and Testing

### e.1 Introduction

Relationships between the target variable – loan status – and the loan applicant features such as gender, civil status, education, etc. were explored through data visualization. This uncovered interesting associations between the variables; however, these associations don't automatically mean interdependence between the variables. These relationships could have just been random, and therefore, meaningless.

Statistical tests were designed to measure the significance of these apparent relationships. Throughout this section, a significance level of 5% ($\alpha = 0.05$) was chosen to signify that interpreted results are true at 95% confidence level. Throughout this section also, the dataset that will be used has been cleaned and the missing fields have been imputed with the appropriate values.

The target variable is a binary categorical with values either Y or N. The rest of the columns are either categorical or continuous variables.

In the case of the relationships between the target variable and categorical variables, provided that the samples being tested conforms to the assumptions of the chi-square test, then the chi-square test of independence will be used for hypothesis testing,

In using the chi-square test, if associations were found, the residuals resulting from the test will be visualized to gain deeper insight within the relationship. The visualization will display the residuals in blue and red gradient of colours. A positive residual will be assigned to a certain shade of blue. The shade will depend on the value of residuals. The same scheme will be applied on the negative residuals except that the colour that will be used is red.

A positive residual means an attraction (positive association) between the corresponding row and column variables, while a negative residual implies a repulsion (negative association) between the corresponding row and column variables [1].

Parametric or non-parametric tests may be used to test the means of two groups from the categorical target Loan_Status with respect to the continuous variables. However, there are several issues to consider as implied by the data visualization, such as the non-normal distribution, imbalanced samples, and high probability of difference in variances. As a solution, the variables were logarithmically transformed as an attempt to make it conform to a normal data distribution. Shapiro-Wilk Method will help assess if normality of distribution is achieved prior to deciding what type of statistical test will give valid results. .

[1] Chi-Square Test of Independence in R - Easy Guides - Wiki - STHDA

The results of these hypothesis tests will be helpful in identifying significant predictors for the machine learning algorithms that will be used to predict loan status later on in this project.

### e.1.1 Types and Assumptions of a Hypothesis Testing

Statistical hypothesis testing requires several assumptions. These assumptions include considerations of the level of measurement of the variable, the method of sampling, the shape of the population distribution, and the sample size. The specific assumptions may vary, depending on the test or the conditions of testing. However, without exception, all statistical tests assume random sampling. Tests of hypotheses about means also assume interval-ratio level of measurement and require that the population under consideration be normally distributed or that the sample size be larger than 50 [2].

#### e.1.1.1 Chi-Square Test and its Assumptions [3]

The Chi-Square Test of Independence determines whether there is an association between categorical variables (i.e., whether the variables are independent or related). It is a nonparametric test.[4]

It can only compare categorical variables. It cannot make comparisons between continuous variables or between categorical and continuous variables. Additionally, the Chi-Square Test of Independence only assesses associations between categorical variables, and can not provide any inferences about causation.

Assumptions are as follows:

- Two categorical variables.
- Two or more categories (groups) for each variable.
- Independence of observations.
    - There is no relationship between the subjects in each group.
    - The categorical variables are not "paired" in any way (e.g. pre-test/post-test observations).
- Relatively large sample size.
    - Expected frequencies for each cell are at least 1.
    - Expected frequencies should be at least 5 for the majority (80%) of the cells.

[2] https://www.sagepub.com/sites/default/files/upm-binaries/43443_7.pdf
[3] The Chi-square test of independence (nih.gov)
4 https://libguides.library.kent.edu/spss/chisquare

### e.1.1.2 t-test and Assumptions

A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features.

A t-test looks at the t-statistic, the t-distribution values, and the degrees of freedom to determine the statistical significance. To conduct a test with three or more means, one must use an analysis of variance.[5]

**Assumptions of t-test statistical testing:**

- Data values must be independent. Measurements for one observation do not affect measurements for any other observation.
- Data in each group must be obtained via a random sample from the population.
- Data in each group is normally distributed.
- Data values are continuous.
- The variances for the two independent groups are equal.

### e.1.1.3 Comparison of Welch t-test[6]

- Welch's t-test, unlike Student's t-test, does not have the assumption of equal variance (however, both tests have the assumption of normality). When two groups have equal sample sizes and variances, Welch's tends to give the same result as Student's. However, when sample sizes and variances are unequal, Student's t-test is quite unreliable; Welch's tends to perform better.

### e.1.1.4 Analysis of Variance

- Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests.[7]
- A one-way ANOVA is commonly used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.It may also be used for only two groups. [8]
- If no true variance exists between the groups, the ANOVA's F-ratio should equal close to 1.

---

[5] Two-Sample t-Test | Introduction to Statistics | JMP

[6] https://www.statisticshowto.com/welchs-test-for-unequal-variances/

7 https://www.investopedia.com/terms/a/anova.asp

8 https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics.php

### e.1.1.5 Kruskal-Wallis

- Kruskal-Wallis test by rank is a non-parametric alternative to one-way ANOVA test, which extends the two-samples Wilcoxon test in the situation where there are more than two groups. It's recommended when the assumptions of one-way ANOVA test are not met. [9]

### e.1.1.6 Correlation tests[10]

A correlation coefficient measures the extent to which two variables tend to change together. The coefficient describes both the strength and the direction of the relationship.Two of the correlation analyses are the following :

**Pearson product moment correlation**

- The Pearson correlation evaluates the linear relationship between two continuous variables. A relationship is linear when a change in one variable is associated with a proportional change in the other variable.
- For example, you might use a Pearson correlation to evaluate whether increases in temperature at your production facility are associated with decreasing thickness of your chocolate coating.

**Spearman rank-order correlation**

- The Spearman correlation evaluates the monotonic relationship between two continuous or ordinal variables. In a monotonic relationship, the variables tend to change together, but not necessarily at a constant rate. The Spearman correlation coefficient is based on the ranked values for each variable rather than the raw data.
- Spearman correlation is often used to evaluate relationships involving ordinal variables. For example, you might use a Spearman correlation to evaluate whether the order in which employees complete a test exercise is related to the number of months they have been employed.

### e.1.1.7 Covariance

- Covariance indicates the relationship of two variables whenever one variable changes. If an increase in one variable results in an increase in the other variable, both variables are said to have a

9 http://www.sthda.com/english/wiki/kruskal-wallis-test-in-r

10 https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/

positive covariance. Decreases in one variable also cause a decrease in the other. Both variables move together in the same direction when they change. Decreases in one variable resulting in the opposite change in the other variable are referred to as negative covariance. These variables are inversely related and always move in different directions.

**e.2 Summary of Hypotheses Formulated:**

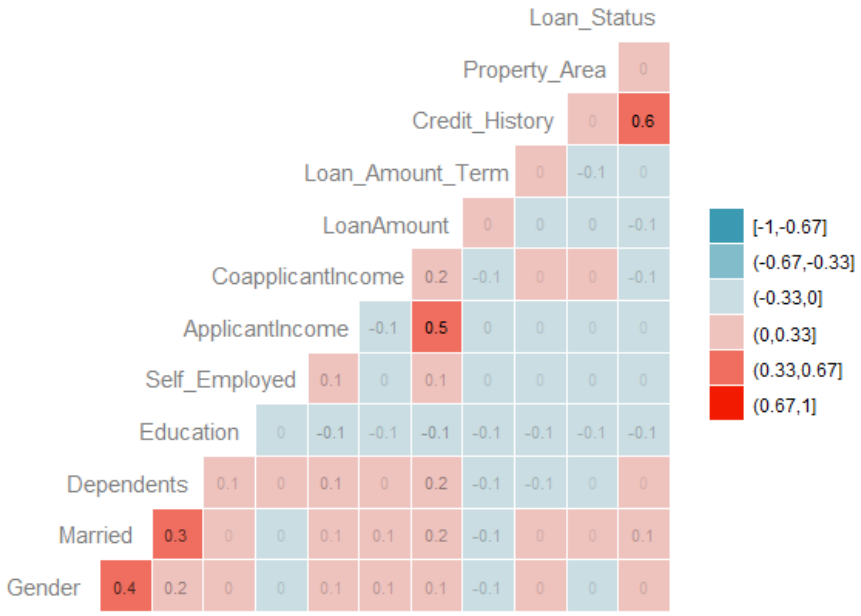
**Table 8 Hypotheses Formulated from Data Visualizations**

The highlighted hypotheses showed consistency of change in Loan_Status with respect to the independent variables.

| | |
|---|---|
| Hypothesis No. 1: | Males are more likely to be eligible than females |
| Hypothesis No. 2: | A married applicant is more likely to be eligible for a house loan |
| Hypothesis No. 3: | An applicant with no dependents is more likely to be eligible for a house loan |
| Hypothesis No. 4: | Graduates are more likely to be eligible for a house loan |
| Hypothesis No. 5: | Self-employed are less likely to be eligible for a house loans |
| Hypothesis No. 6: | Applicants who met the credit history guidelines are more likely to be eligible for a house loan |
| Hypothesis No. 7: | Applicants who apply for Semiurban area are more likely to be eligible for a house loan. |
| Hypothesis No. 8: | Applicants who earn more are more likely to be eligible. |
| Hypothesis No. 9: | Applicants who apply for less loan amount are more likely to be eligible. |
| Hypothesis No. 10: | Applicants who can pay sooner are more eligible. |
| Hypothesis No. 11: | Applicants with lower EMI are more likely to be eligible |
| Hypothesis No. 12: | Applicants with lower Debt to Income Ratio are more likely to be eligible |

## f. Hypothesis Testing

We can have a correlation matrix of all variables. From here we can make initial inferences with respect to all the hypotheses made. These inferences are not final, but are useful to filter which variables show consistent positive or negative linear relationship to Loan_Status.




**Interpretation of Correlation Matrix against Loan_Status :**

**Reference:** +1 or -1 are the perfect correlation coefficients. +1 implies positive linear relationship and -1 implies perfect negative linear relationship.

**Positive**

- **Married (+) weak positive** correlation (+0.1)
  - Married applicants may have higher likelihood of being eligible
- **Credit History (+) strongest positive** correlation (+0.6)
  - Applicants whose credit history passed the guidelines may have a higher likelihood of being eligible than those who did not pass the guidelines.
- **Dependents, Gender and Property_Area (+) weak positive** correlation
  - Males may have a higher likelihood of being eligible than females.
  - Applicants applying for semi-urban and/or urban areas may have a higher likelihood of being eligible compared to those in rural areas.

**Negative and 0**

- **Education (-) weak negative** correlation (-0.1)
  - More educated applicants may have higher likelihood to be eligible.
- **Loan_Amount (-) negative** correlation

    - Applicants applying for relatively lower loan amounts may have a higher chance of being eligible.
- **Coapplicant_Income (-) weak negative** correlation
    - As coapplicant income decreases, the likelihood of being eligible increases.
- **ApplicantIncome (0) correlation**

**Independent Variable correlations:**

There are several predictor variables showing positive linear relationships with each other. This is important in pruning down predictors for data modelings.

- Applicant Income and Loan Income have strong positive linear relationship.
- Coapplicant Income and Applicant Income have weak negative linear relationship.
- Gender has weak positive linear relationships with Married, Incomes, Loan Amount and Credit History.
- Education has weak negative linear relationships with Incomes, Loan Amount, Term, Credit History and Property area.
- Married has strong positive linear relationships with Gender, and Dependents and weak positive linear relationships with Incomes, Loan_Amount. and weak negative linear relationship with Loan Term.

**Code:**

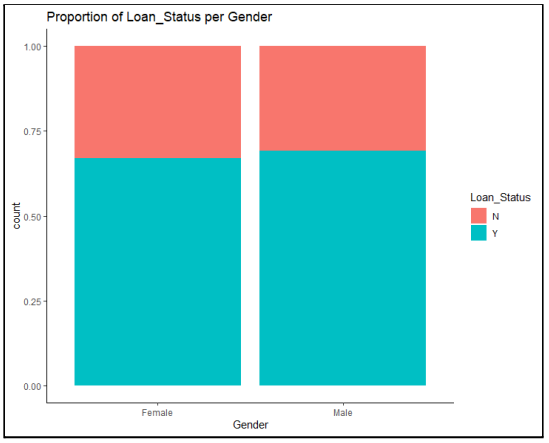
```
#Correlation Matrix of all variables
library(GGally)
# Convert data to numeric
corr <- data.frame(lapply(hloan[-1], as.numeric))
corr
# Plot the graph
ggcorr(corr,
       method = c("pairwise", "pearson"),
       nbreaks = 6,
       hjust = 0.9,
       label = TRUE,
       label_size = 3,
       label_alpha = TRUE,
       color = "grey50",
       layout.exp = 1)
```

**Hypothesis No. 1: Males are more likely to be eligible than females**

| | |
|---|---|
| **Ho:** | There is no significant relationship between gender and loan status. |
| **Ha:** | There is a significant relationship between gender and loan status. |

The data consists of two categorical variables – gender and loan status. A chi-square test is the appropriate method of testing the null hypothesis.

The following code will create a table of frequency of Gender with respect to the Loan_Status:




**Code:**

```
table(ins_transformed$Gender,ins_transformed$Loan_Status)
```

**Output:**

| | N | Y |
|---|---|---|
| Female | 39 | 76 |
| Male | 153 | 346 |

**Chi-square test:**

**Code:**

```
#Chi-square test of Gender and Loan_Status
gender <- chisq.test(table(ins_transformed$Gender,
ins_transformed$Loan_Status),correct = F)gender
```

**Output:**

```
        Pearson's Chi-squared test

data:  table(ins_imputed$Gender, ins_imputed$Loan_Status)
X-squared = 0.45981, df = 1, p-value = 0.4977
```

**Conclusion by interpretation of p-value:**

Since the p-value is greater than our chosen significance level (α = 0.05), we do not reject the null hypothesis. Rather, we conclude that there is not enough

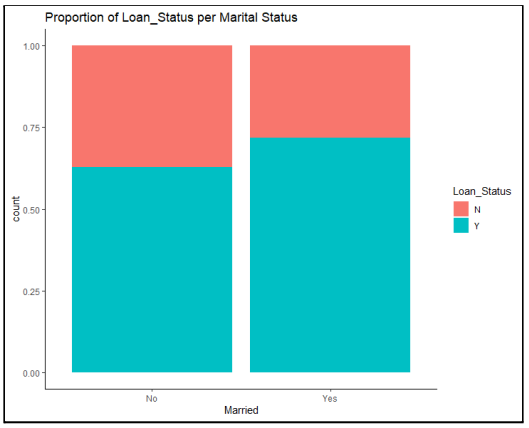
| |
|---|
| evidence to suggest a relationship between gender and loan status. |

**Hypothesis No. 2: A married applicant is more likely to be eligible for a house loan**

| | |
|---|---|
| **Ho:** | There is no significant relationship between civil status and eligibility. |
| **Ha:** | There is a significant relationship between civil status and eligibility. |

The data consists of two categorical variables – married and loan status. A chi-square test is the appropriate method of testing the null hypothesis.

The following R script will create a table of Married frequency with respect to Loan_Status.




**Code:**

```
table(ins_transformed$Married,ins_transformed$Loan_Status)
```

**Output:**

```
     N   Y
No   79 135
Yes 113 287
```

**Chi-square test:**

**Code:**

```
#Chi-square test of Married and Loan_Status
chisq.test(table(ins_transformed$Married,
ins_transformed$Loan_Status),correct = F)
```

**Output:**

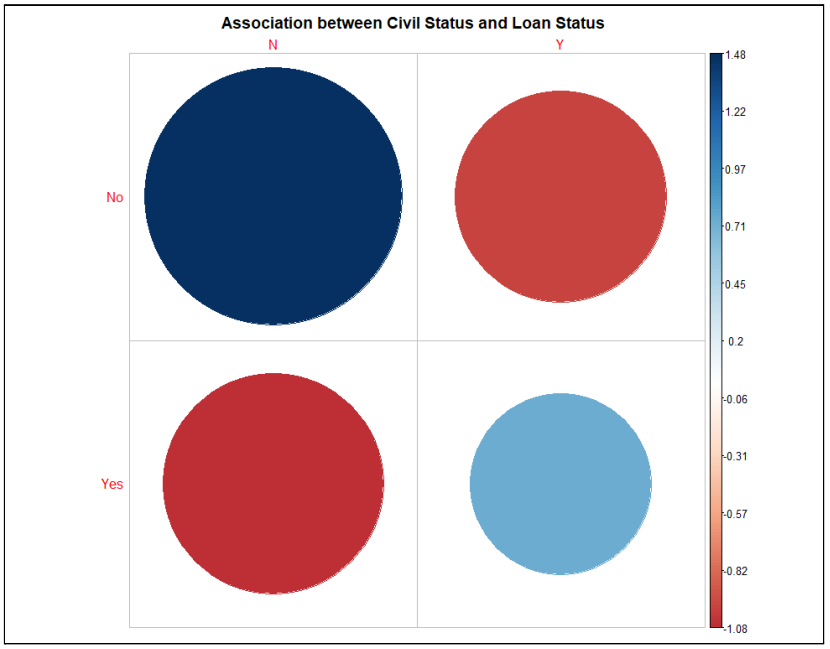
```
        Pearson's Chi-squared test

data:  table(ins_transformed$Married, ins_transformed$Loan_Status)
X-squared = 4.8714, df = 1, p-value = 0.02731
```

**Conclusion by interpretation of p-value:**

Since $p - value = 0.02731 < alpha = 0.05$, we can reject the null hypothesis, and conclude that there is a significant relationship between applicant civil status and loan status.

**Verify by correlation:**

The conclusion above is further supported by correlation test results.




**Interpretation:**

1. Married applicants are strongly and positively associated with loan eligibility.
2. Unmarried applicants are strongly and negatively associated with loan eligibility.

**Code:**

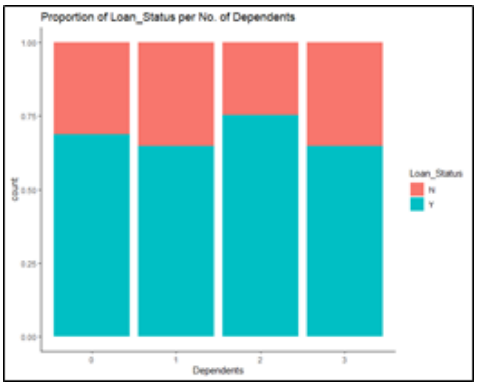
```
#Correlation plot of residuals of Married and Loan_Status
library(corrplot)
corrplot(married$residuals,is.corr = F,
         title = "Association between Civil Status and Loan Status",
         mar=c(0,0,2,0))
```

**Hypothesis No. 3: An applicant with no dependents is more likely to be eligible for a house loan**

| | |
|---|---|
| **Ho:** | There is no significant relationship between number of dependents and eligibility. |
| **Ha:** | There is a significant relationship between number of dependents and eligibility. |

The data consists of two categorical variables – number of dependents and loan status. A chi-square test is the appropriate method of testing the null hypothesis.

The following R script will create a table of frequency of Number of Dependents with respect to Loan_Status.




**Code**:
```
table(ins_transformed$Dependents,ins_transformed$Loan_Status)
```

**Output:**

| | N | Y |
|---|---|---|
| 0 | 109 | 243 |
| 1 | 36 | 69 |
| 2 | 28 | 77 |
| 3 | 19 | 33 |

**Chi-square test:**

**Code:**
```
#Chi-square test of Dependents and Loan_Status
dependents <- chisq.test(table(ins_transformed$Dependents,
ins_transformed$Loan_Status),correct = F)
```

**Output:**

```
        Pearson's Chi-squared test

data:  table(ins_transformed$Dependents, ins_transformed$Loan_Status)
X-squared = 2.1663, df = 3, p-value = 0.5386
```

**Conclusion by interpretation of p-value:**

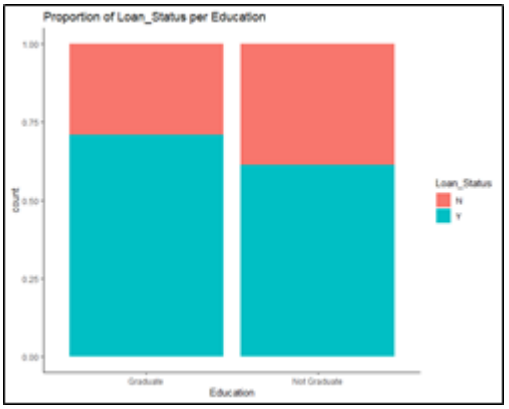
Since the$p - value = 0.5386 > alpha = 0.05$, we failed to reject the null hypothesis. Rather, we conclude that there is not enough evidence to suggest a relationship between number of dependents and loan status.

**Hypothesis No. 4: Graduates are more likely to be eligible for a house loan**

| | |
|---|---|
| **Ho:** | There is no significant relationship between education and eligibility. |
| **Ha:** | There is a significant relationship between education and eligibility. |

The data consists of two categorical variables – education and loan status. A chi-square test is the appropriate method of testing the null hypothesis.

The following R script will create a frequency table of Education with respect to Loan_Status:




**Code:**

```
table(ins_transformed$Education,ins_transformed$Loan_Status)
```

**Output:**

```
              N   Y
Graduate    140 340
Not Graduate  52  82
```

**Chi-square test:**

**Code:**

```
#Chi-square test of Education and Loan_Status
education <- chisq.test(table(ins_transformed$Education,
             ins_transformed$Loan_Status),correct = F)
```

**Output:**

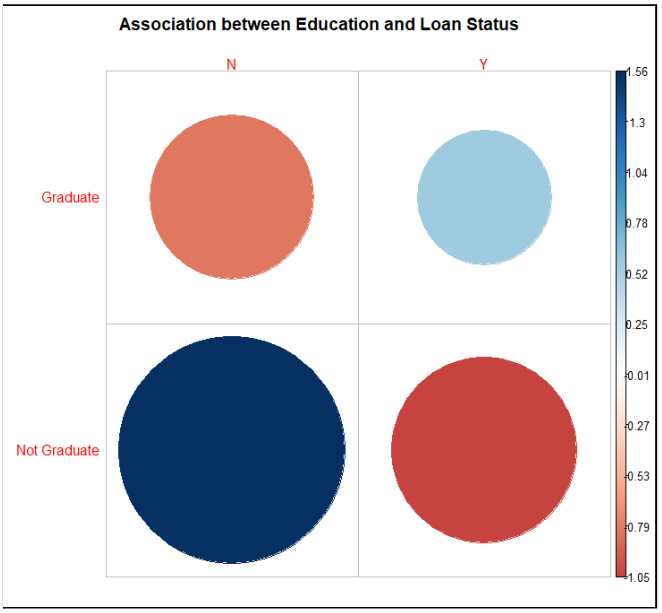
```
        Pearson's Chi-squared test

data:  table(ins_transformed$Education, ins_transformed$Loan_Status)
X-squared = 4.5289, df = 1, p-value = 0.03333
```

**Conclusion by interpretation of p-value:**

Since the $p - value = 0.03333 < alpha = 0.05$, we can reject the null hypothesis, and conclude that there is a significant relationship between educational background and loan status.

**Verify by correlation:**

The conclusion above is further supported by correlating the residuals of Education and Loan_Status.




**Interpretation:**

1. Educated applicants are positively associated with loan eligibility.
2. Applicants who did not graduate are strongly and negatively associated with loan eligibility.

**Code:**

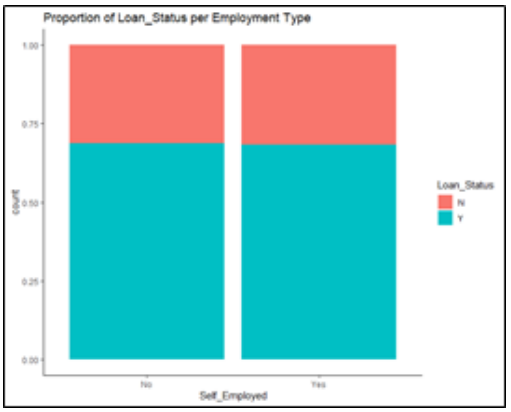
```
#Correlation plot of residuals of Education and Loan Status
corrplot(education$residuals,is.corr = F,
         title = "Association between Education and Loan Status",
         mar=c(0,0,2,0),tl.srt=0)
```

**Hypothesis No. 5: Self-employed are less likely to be eligible for a house loan**

| | |
|---|---|
| **Ho:** | There is no significant relationship between employment type and eligibility. |
| **Ha:** | There is a significant relationship between employment type and eligibility. |

The data consists of two categorical variables – self-employment and loan status. A chi-square test is the appropriate method of testing the null hypothesis.

To get the frequencies of 2(two) variables being tested, run this code in R




**Code:**

```
table(ins_transformed$Self_Employed,ins_transformed$Loan_Status)
```

**Output:**

| | N | Y |
|---|---|---|
| No | 164 | 365 |
| Yes | 28 | 57 |

**Chi-square test:**

**Code:**

```
#Chi-square test of Self_Employed and Loan_Status
chisq.test(table(ins_transformed$Self_Employed,
ins_transformed$Loan_Status),correct = F)
```

**Output:**

```
        Pearson's Chi-squared test

data:  table(ins_transformed$Self_Employed, ins_transformed$Loan_Status)
X-squared = 0.12815, df = 1, p-value = 0.7204
```

**Conclusion by interpretation of p-value:**

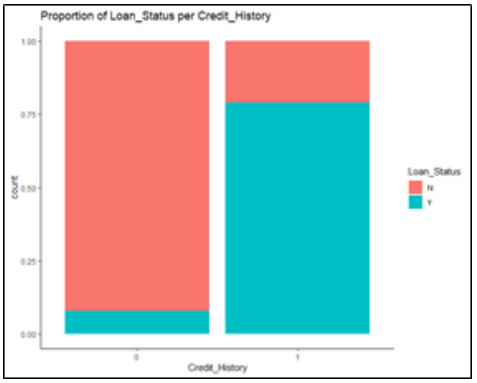
Since $p - value = 0.7204 > alpha = 0.05$, we failed to reject the null hypothesis. Rather, we conclude that there is not enough evidence to suggest a relationship between self-employment and loan status.

**Hypothesis No. 6: Applicants who met the credit history guidelines are more likely to be eligible for a house loan**

| | |
|---|---|
| **Ho:** | There is no significant relationship between credit history and eligibility. |
| **Ha:** | There is a significant relationship between credit history and eligibility. |

The data consists of two categorical variables – credit history and loan status. A chi-square test is the appropriate method of testing the null hypothesis.

The following R script will create a frequency table of Credit_History with respect to Loan_Status:




**Code:**

```
table(ins_transformed$Credit_History,ins_transformed$Loan_Status)
```

**Output:**

| | N | Y |
|---|---|---|
| 0 | 90 | 9 |
| 1 | 102 | 413 |

**Chi-square test:**

**Code:**

```
#Chi-square test of Credit_History and Loan Status
credit <- chisq.test(table(ins_transformed$Credit_History,
ins_transformed$Loan_Status),correct = F)
```

**Output:**

```
        Pearson's Chi-squared test

data:  table(ins_transformed$Credit_History, ins_transformed$Loan_Status)
X-squared = 195.33, df = 1, p-value < 2.2e-16
```

**Conclusion by interpretation of p-value:**

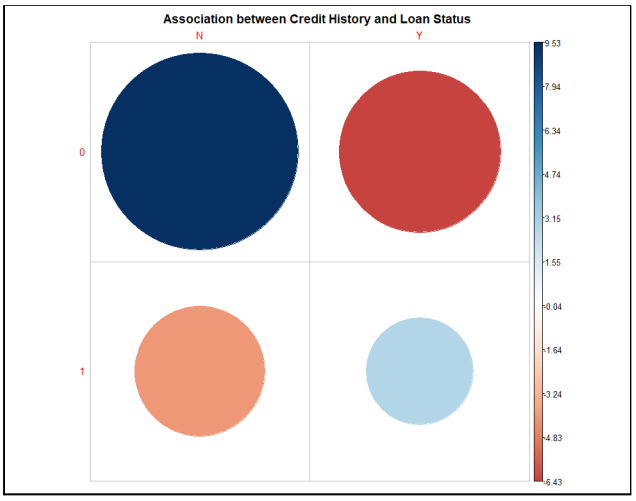
Since $p - value = 2.179098e - 44 < alpha = 0.05$, we can reject the null hypothesis, and conclude that there is a significant relationship between credit history and loan status.

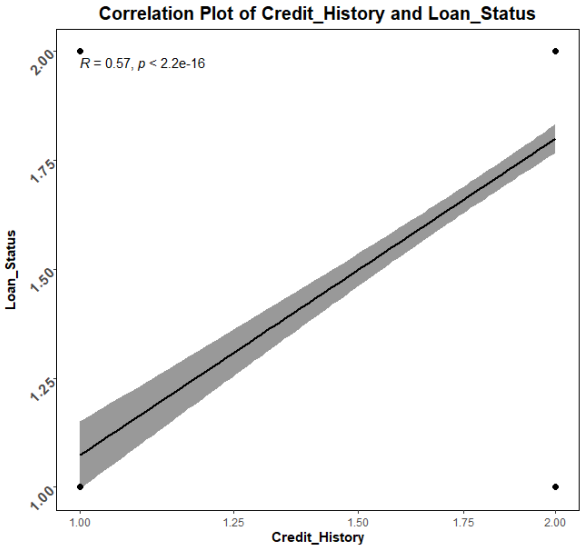
**Verify by correlation:**

1. Having not met the guidelines in credit history is strongly associated with loan ineligibility.
2. Having met the guidelines in credit history is associated with loan eligibility.

**Code:**
*corrplot(credit$residulas, is.corr=F, title="Association between Credit History and Loan Status,", mar=c(0,2,0), tl.srt=0)*

**Output:**



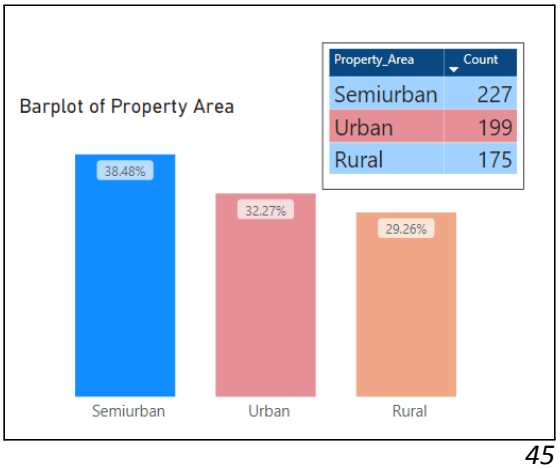




**Hypothesis No. 7: Applicants who apply for Semiurban area are more likely to be eligible for a house loan.**

| | |
|---|---|
| **Ho:** | There is no significant relationship between property area and eligibility. |
| **Ha:** | There is a significant relationship between property area and eligibility. |

The data consists of two categorical variables – property area and loan status. A chi-square test is the appropriate method of testing the null hypothesis.

The following R script will execute the frequency of Property_Area relative to Loan_Status:



**Code:**

```
table(ins_transformed$Property_Area,ins_transformed$Loan_Status
```

**Output:**

| | N | Y |
|---|---|---|
| Rural | 69 | 110 |
| Semiurban | 54 | 179 |
| Urban | 69 | 133 |

**Chi-square test:**

**Code:**

```
#Chi-square test of Property Area and Loan Status
property <- chisq.test(table(ins_transformed$Property_Area,
ins_transformed$Loan_Status),correct = F
```

**Output:**

```
        Pearson's Chi-squared test

data:  table(ins_transformed$Property_Area, ins_transformed$Loan_Status)
X-squared = 12.298, df = 2, p-value = 0.002136
```

**Conclusion by interpretation of p-value:**

Since the$p - value = 0.002136 < alpha = 0.05$, we can reject the null hypothesis, and conclude that there is a significant relationship between property area and loan status.

**Verify by correlation:**

**Code:**

```
#Correlation plot of Property Residuals and Loan Status
corrplot(property$residuals,is.corr = F,
         title = "Association between Property Area and Loan
Status",
         mar=c(0,0,2,0),tl.srt=0)
```

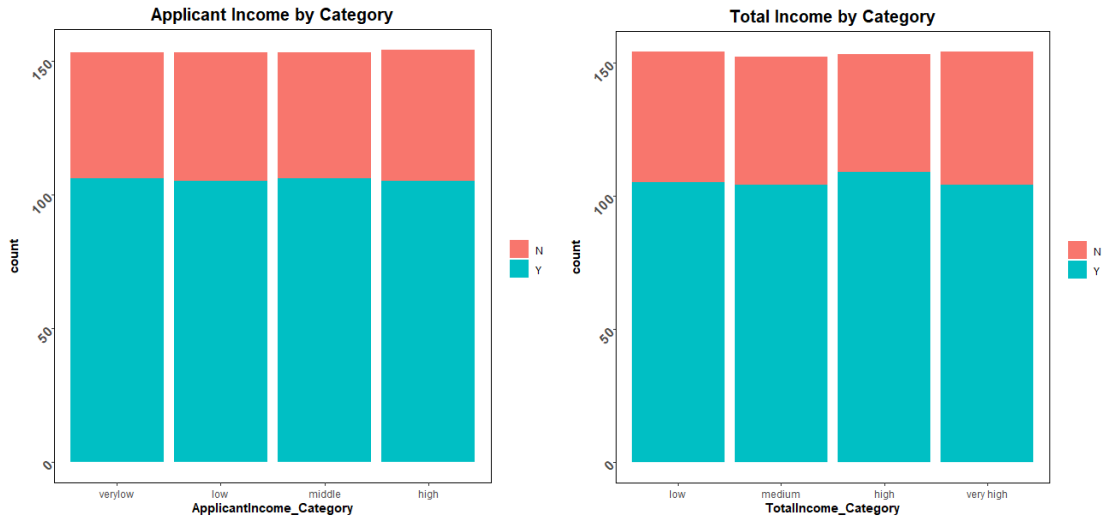
**Interpretation:**

1. Semiurban property areas have the strongest positive association with loan eligibility.

2. Rural property areas have the strongest negative association with loan eligibility.
3. Urban property areas have weak negative association with loan eligibility.
4. There is a strong negative association between semiurban property area and loan ineligibility.

**Hypothesis No. 8: Applicants who earn more are more likely to be eligible.**

| | |
|---|---|
| **Ho:** | There is no significant relationship between applicant income and eligibility. |
| **Ha:** | There is a significant relationship between applicant income and eligibility. |




Using feature engineering, the applicant income and total income are both divided into 4 categories. From here, we can make an early inference that Loan eligibility is not affected by income. This will be validated by statistical tests below.

**Code:**

**Code for feature engineering:**

```
#Separate applicant income into 4 quantile bins
ai <- hloan %>% mutate(ApplicantIncome_Category=cut(ApplicantIncome,
breaks=unique((quantile(ApplicantIncome, probs=seq.int(0,1,
by=1/4)))), labels=c("verylow","low","middle","high")))

#Plot
filter(ai, !is.na(ApplicantIncome_Category)) %>%  ggplot() +
  geom_bar(mapping = aes(x = ApplicantIncome_Category, fill =
Loan_Status)) +
  scale_fill_discrete(na.translate=FALSE)+mytheme+
  ggtitle("Applicant Income by Category")
```

**Look for the appropriate statistics test.**

**Three options: Anova, Krugal-Wallis Test and Chi-square**

### Test 1: Use Anova on log transformed ApplicantIncome and Total Income

**Code:**

```
#Use Anova on log transformed ApplicantIncome and Total Income
res.aov <- aov(ApplicantIncome_log ~ Loan_Status, data = hloan)
res.aov <- aov(TotalIncome_log ~ Loan_Status, data = hloan)
```

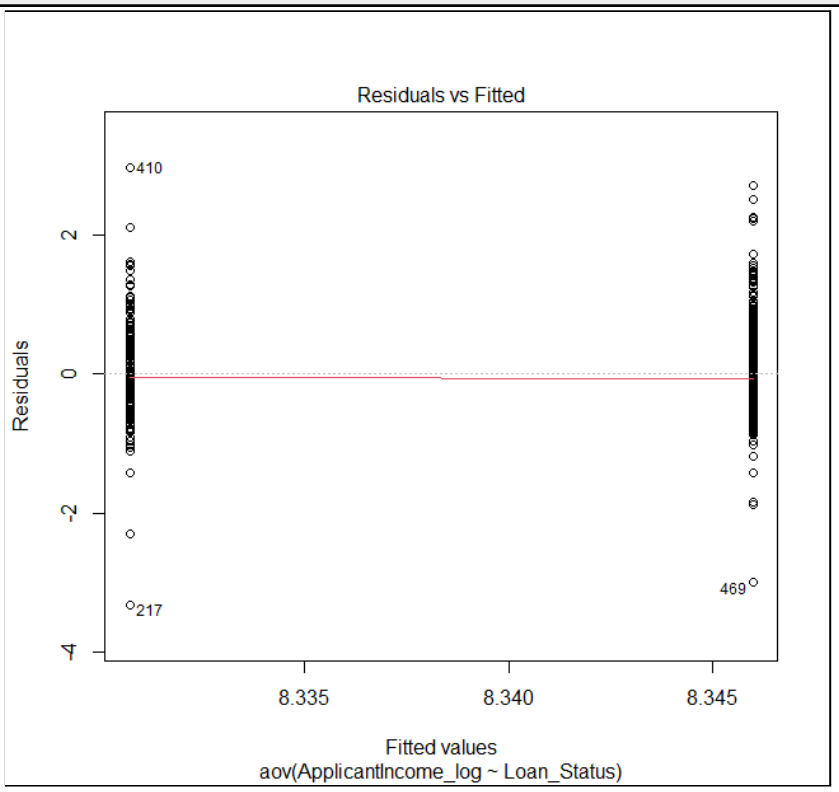
```
Call:
   aov(formula = ApplicantIncome_log ~ Loan_Status, data = hloan)

Terms:
                Loan_Status Residuals
Sum of Squares      0.03075 255.20075
Deg. of Freedom           1       612

Residual standard error: 0.6457513
Estimated effects may be unbalanced
```

### Check validity of Anova Result

Normal population should have homogeneous variances. To plot it:

**Code:**

```
# check validity of anova result
# check homogeneity of variances
plot(res.aov, 1)
```




A statistical test for the homogeneity of variance in this assumed normal and transformed population is the Levene Test.

**Code:**

```
#test homogeneity of variance
leveneTest(ApplicantIncome_log ~ Loan_Status, data = hloan)
```

**Output:**

```
> leveneTest(ApplicantIncome_log ~ Loan_Status, data = hloan)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   1  0.3203 0.5716
      612
```

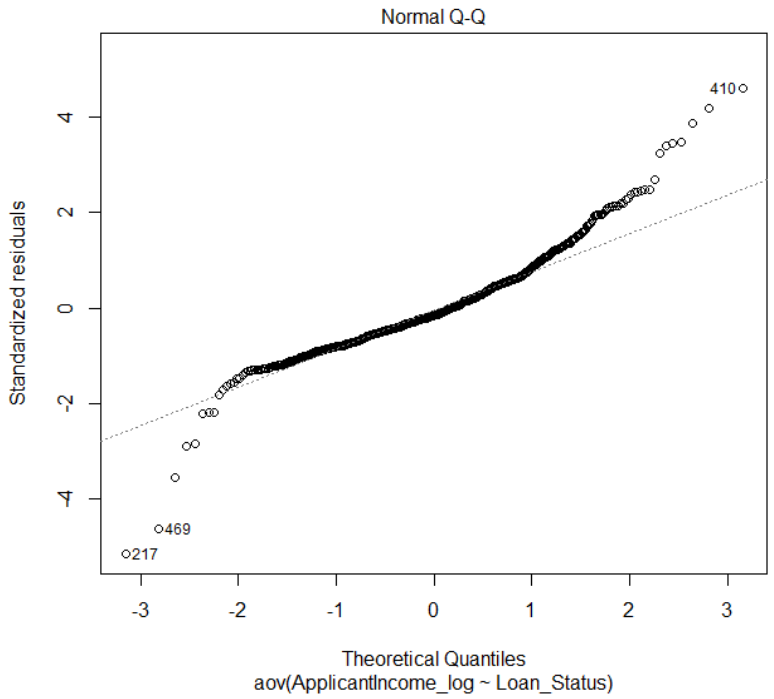
**Homogeneity of variance:**

Since p value = 0.5716 > alpha =0.05, then this means that there is no evidence to suggest that the variance of Income across Loan_Status categories is statistically significantly different. Therefore, we can assume the homogeneity of variances of Applicant Income in Loan_Status categories.

**Check assumption of normality:**

**Code:**

```
#check normality of residuals
plot(res.aov, 2)
```




In the plot above, the quantiles of the residuals are plotted against the quantiles of the normal distribution. A 45-degree reference line is also plotted.

The normal probability plot of residuals is used to check the assumption that the residuals are normally distributed. It should approximately follow a straight line. It is observable in the plot that while the majority of residuals follow a straight line, a good portion also go above the line, with some outliers also going below.

This observation is supported by the Shapiro-Wilk test on the Anova residuals, that normality is violated.

**Code:**

```
# Extract the residuals
aov_residuals <- residuals(object = res.aov )
# Run Shapiro-Wilk test
shapiro.test(x = aov_residuals )
```

**Output:**

```
        Shapiro-Wilk normality test

data:  aov_residuals
W = 0.93529, p-value = 1.164e-15
```

**Failed assumption of normality:**

Since the normality assumption of Anova is violated, Kruskal Test will give a more reliable statistical test result.

**Test no. 2**: **Kruskal Test**

The means of income (ApplicantIncome_log, and TotalIncome_log) of the two independent groups (Loan_Status = Y and Loan_Status = N) will be compared via Kruskal Test.

**Code:**

```
kruskal.test(ApplicantIncome_log ~ Loan_Status, data = hloan)
```

**Output:**

```
        Kruskal-Wallis rank sum test

data:  ApplicantIncome_log by Loan_Status
Kruskal-Wallis chi-squared = 0.01062, df = 1, p-value = 0.9179
```

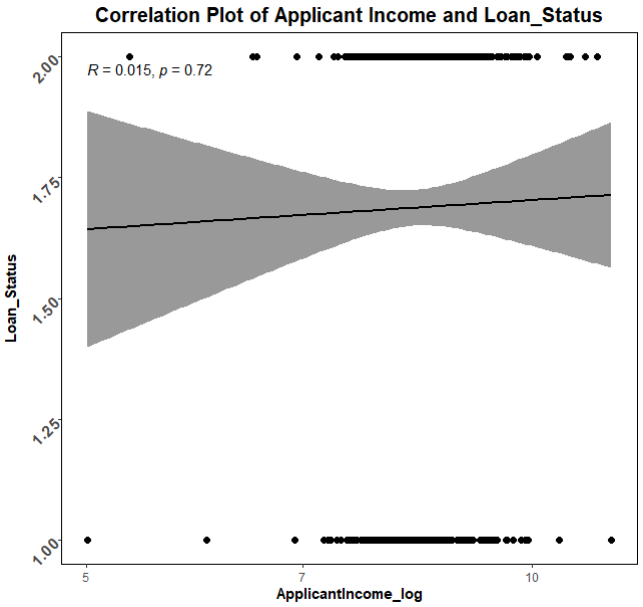
**Conclusion by interpretation of p-value:**

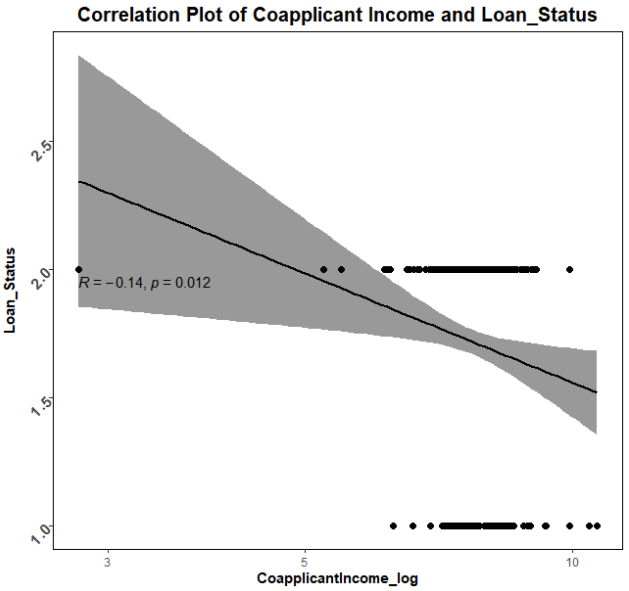
Since p-value = 0.9179 > 0.05, then there is no significant evidence to suggest that the Applicant Income across Loan_Status levels is statistically significantly different, nor does its value directly affect eligibility.

Same inference resulted in Total Income.

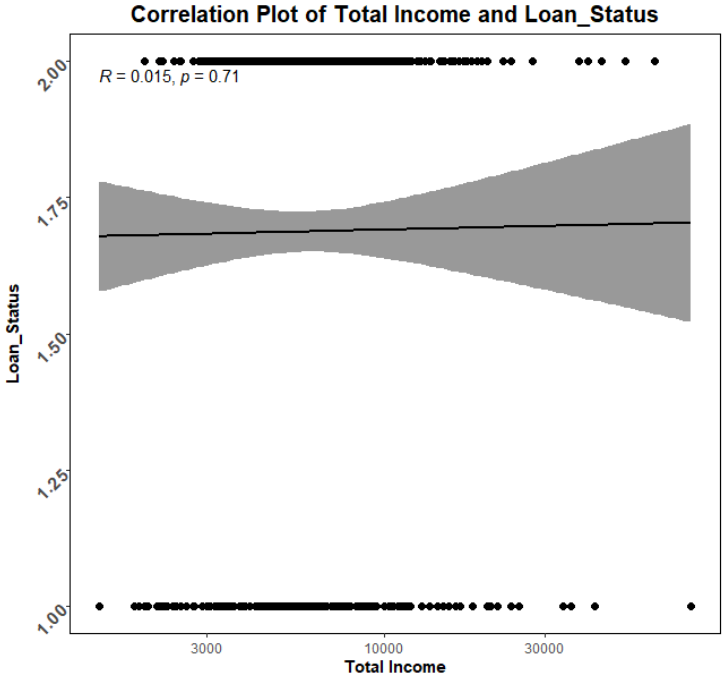
**Verify by correlation:**

Applicant Income and Total: A low rho value of 0.015 denotes a weak positive correlation to loan_status.



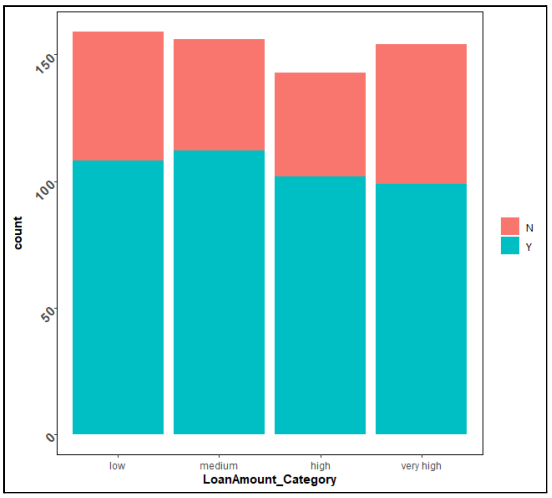



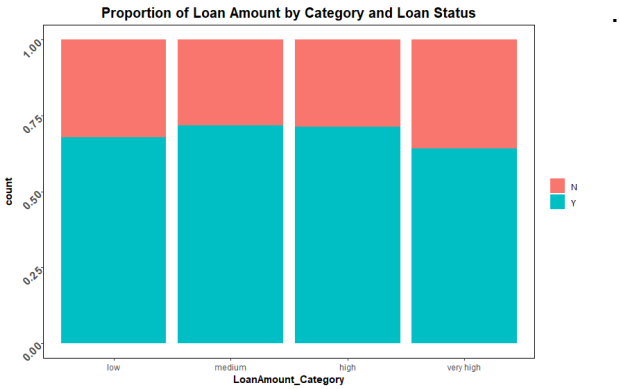

Coapplicant Income: A low rho value of -0.14 denotes a low negative correlation to loan status. Using Kruskal-Wallis test to test for p-value = 0.3867, also denotes that there is no significant evidence to assume that Coapplicant Income has a relationship with Loan_Status.



**Hypothesis No. 9: Applicants who apply for less loan amount are more likely to be eligible.**

| | |
|---|---|
| **Ho:** | There is no significant relationship between loan amount value and eligibility. |
| **Ha:** | There is a significant relationship between loan amount value and eligibility. |




Feature Engineering is applied to divide loan amount into 4 categories.

From the figures above, we can have an early inference that loan amount does not significantly affect eligibility. This can be further supported by statistical tests.

**Type of Test: Krugal-Wallis Test**

Following the same steps as in Income:

**Code:**

```
# Compare means of Loan Amount with respect to Loan_Status
kruskal.test(LoanAmount_log ~ Loan_Status, data = hloan)
```

**Output:**

```
        Kruskal-Wallis rank sum test

data:  LoanAmount_log by Loan_Status
Kruskal-Wallis chi-squared = 0.2888, df = 1, p-value = 0.591
```

**Conclusion by p-value:**

Since $p - value = 0.591 > \alpha = 0.05$, then we fail to reject the null hypothesis, and conclude that there is no significant relationship between Loan Amount and Loan_Status.

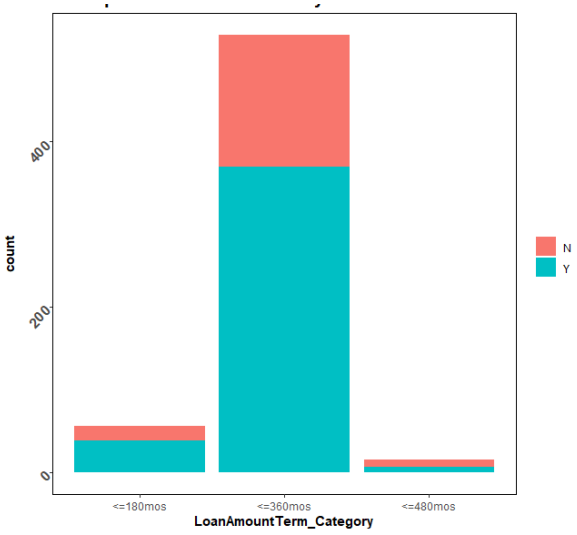
**Confirm by correlation:**

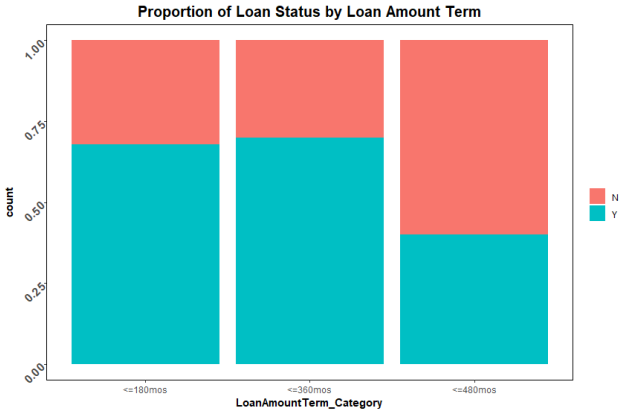
```
        Spearman's rank correlation rho

data:  hloan$LoanAmount_log and as.numeric(hloan$Loan_Status)
S = 39416528, p-value = 0.5914
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.02170532
```

A rho of -0.0217 denotes a low negative correlation of Loan Amount to Loan_Status.

**Hypothesis No. 10: Applicants who can pay sooner are more eligible.**

| | |
|---|---|
| **Ho:** | There is no significant relationship between Loan_Amount_Term and eligibility. |
| **Ha:** | There is a significant relationship between Loan_Amount_Term and eligibility. |




Feature engineering is implemented to divide Loan_Amount_Term into 3 categories.

The data consists of two categorical variables – Loan_Amount_Term and Loan_Status.A chi-square test is the appropriate method of testing the null hypothesis.

Subsets of the data were created to extract the top 3 highest count of loan amount term. This is needed because a chi-square test requires a minimum number of samples. The top 3 loan amount terms in term of count is 360, 180 and 480 months.

**Code:**

```
ins_term <- ins_transformed[ins_transformed$Loan_Amount_Term == 360|
        ins_transformed$Loan_Amount_Term == 180|
        ins_transformed$Loan_Amount_Term == 480,]
```

Ensure that loan amount term is numeric:

**Code:**

```
ins_term$Loan_Amount_Term <- as.numeric(ins_term$Loan_Amount_Terr
```

Derive frequencies of Loan_Amount_Term with respect to Loan_Status category

**Code:**

```
table(ins_term$Loan_Amount_Term, ins_term$Loan_Status)
```

**Output:**

```
      N   Y
180  16  31
360 158 364
480   9   7
```

The following code will execute the chi-square test:

**Code:**

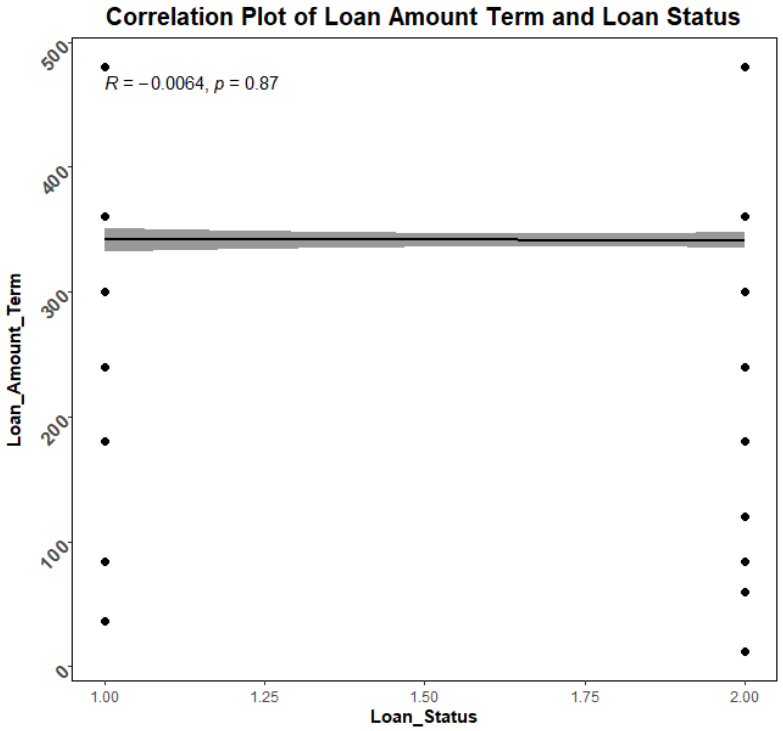
```
chisq.test(table(ins_term$Loan_Amount_Term, ins_term$Loan_Status))
```

```
        Pearson's Chi-squared test

data:  table(ins_term$Loan_Amount_Term, ins_term$Loan_Status)
X-squared = 5.0562, df = 2, p-value = 0.07981
```

**Conclusion by p-value:**

Since $p - value = 0.1795 > alpha = 0.05$, we fail to reject the null hypothesis. Rather, we conclude that there is not enough evidence to suggest a relationship between loan amount term and loan status.

**Confirm by correlation:**



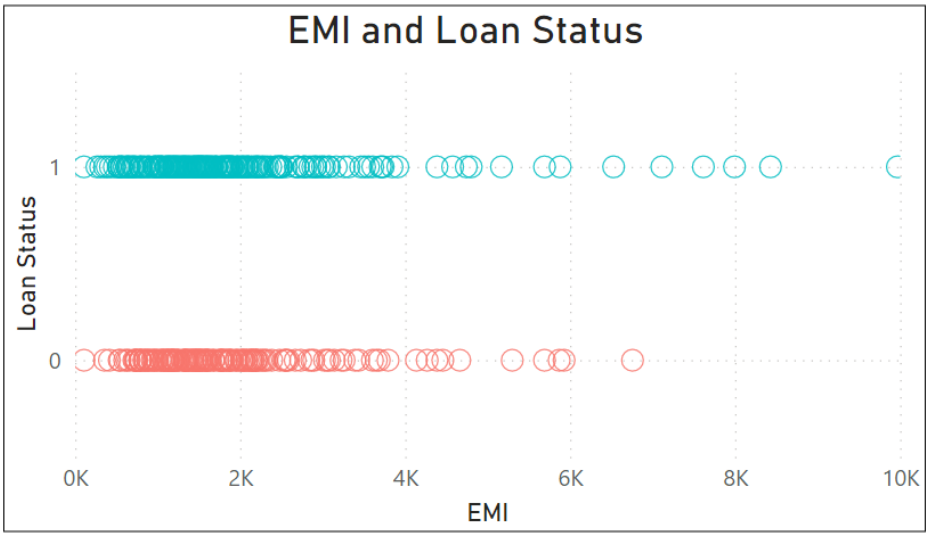

A rho value of -0.006 denotes a low negative linear relationship between Loan_Amount_Term and Loan_Status.

**Code:**

```
#Correlation plot of Loan Amount Term and Loan Status
library("ggpubr")
ggscatter(corr, x = "Loan_Status", y = "Loan_Amount_Term",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "spearman",
          xlab = "Loan_Status", ylab = "Loan_Amount_Term")+mytheme+
          ggtitle("Correlation Plot of Loan Amount Term and Loan
Status")
```

**Hypothesis No. 11: Applicants with lower EMI are more likely to be eligible**

| | |
|---|---|
| **Ho:** | There is no significant relationship between EMI and eligibility. |
| **Ha:** | There is a significant relationship between EMI and eligibility. |




**Statistical Test: chi.sq test**

```
        Pearson's Chi-squared test

data:  table(hloan$EMI, hloan$Loan_Status)
X-squared = 285.39, df = 268, p-value = 0.2223
```

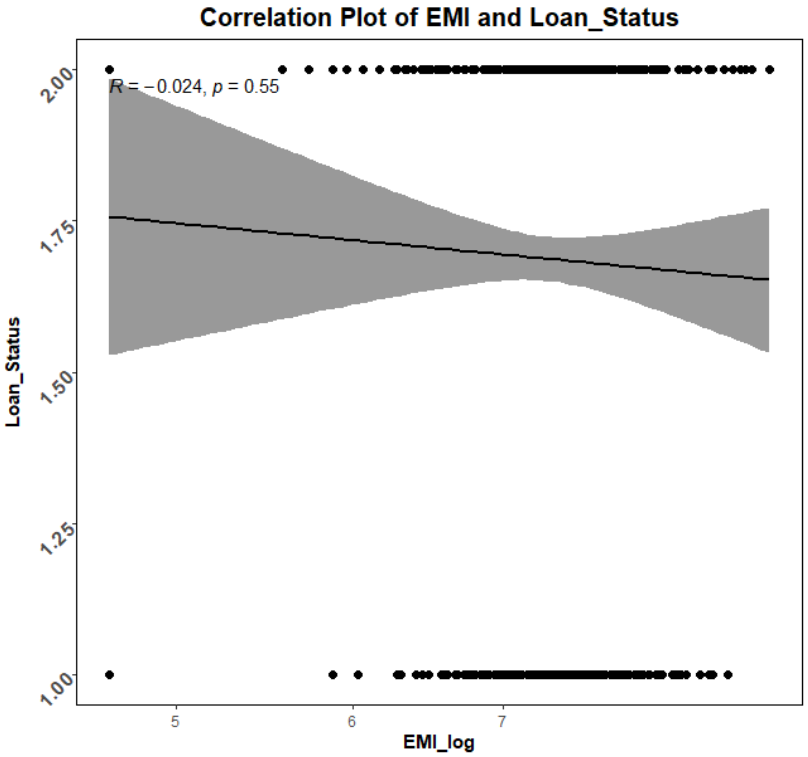
**Conclusion by p-value:**

Since $p - value = 0.2223 > alpha = 0.05$, we failed to reject the null hypothesis and conclude that there is no significant evidence to suggest a relationship between EMI and Loan Status.

**Verify by correlation:**

Correlation test will be used to verify the relationship between EMI and loan status.

The correlation coefficient rho value of $-0.024$ is low. This indicates weak negative correlation with Loan Status.



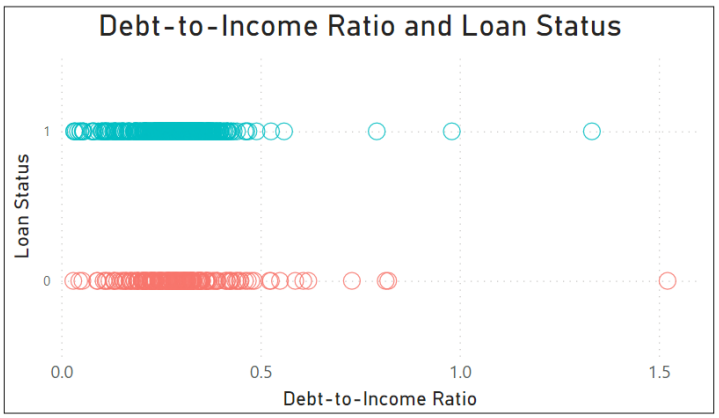

Same results will show by using the code below.

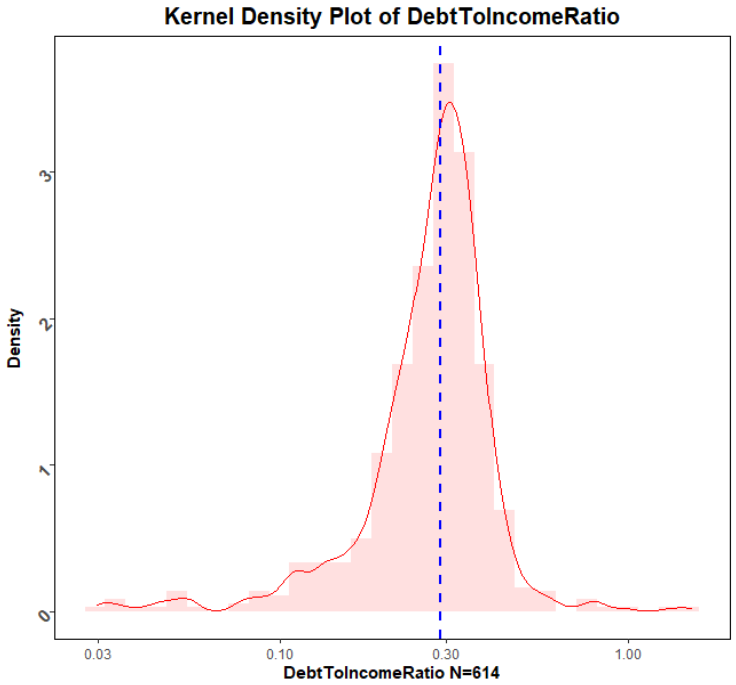
**Code:**

```
#Correlation test between EMI and Loan Status
cor.test(y=hloan$Loan_Status,x=hloan$EMI, method = "spearman", exact
= F)
```

**Hypothesis No. 12: Applicants with lower Debt to Income Ratio are more likely to be eligible**

| | |
|---|---|
| **Ho:** | There is no significant relationship between Debt to Income Ratio and eligibility. |
| **Ha:** | There is a significant relationship between Debt to Income Ratio and eligibility. |




Log transformed Debt to Income Ratio appears normal

**Normality test by Shapiro_Wilk method: Failed**

```
        Shapiro-Wilk normality test

data:  aov_residuals
W = 0.93529, p-value = 1.164e-15
```

A p-value < 0.05 implies that the distribution even after log transformation is not normal. For this reason, Kruskal Willis will be used for statistical test.

**Statistical Test: Kruskal_Willis**

**Code:**

```
#Test the means of Debt to Income Ratio per Loan Status category
kruskal.test(DebtToIncomeRatio_log ~ Loan_Status, data = hloan)
```

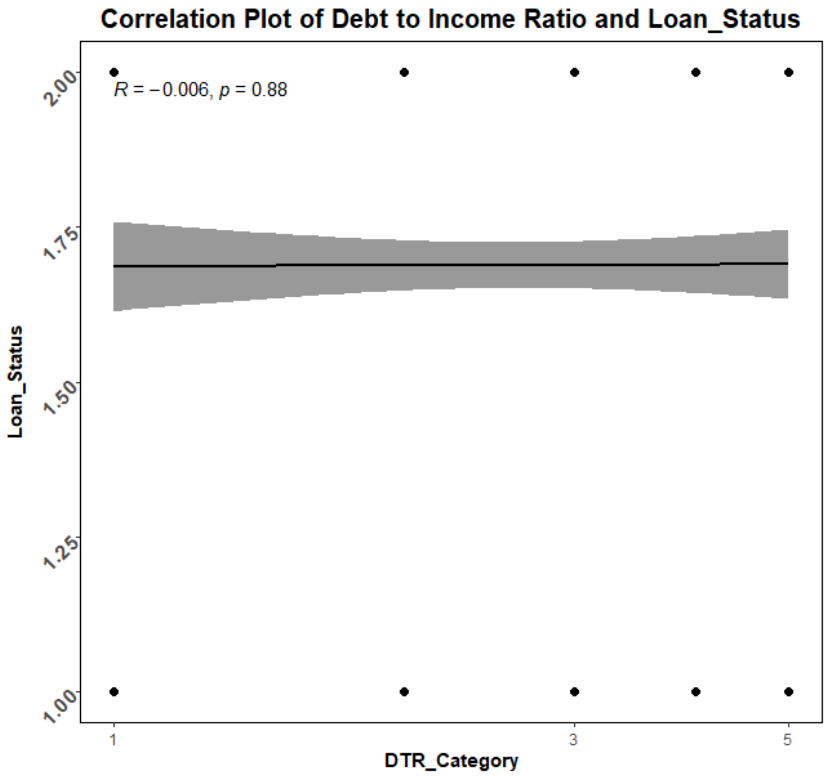
**Conclusion by p-value:**

Since $p - value = 0.818 > alpha = 0.05$, we failed to reject the null hypothesis and conclude that there is no significant evidence to suggest a relationship between Debt to Income Ratio and Loan Status.

**Verify by correlation**

As a confirmation, correlation test using the spearman method is used to test the hypothesis.

The correlation coefficient rho value of $-0.006$ is low. This indicates weak negative correlation to Loan Status.




**Code:**

```
#Correlation test between Debt to Income Ratio and Loan_Status
cor.test(y=hloan$Loan_Status,x=hloan$DebtToIncomeRatio, method =
"spearman",exact = F)
```

**Output:**

```
        Spearman's rank correlation rho

data:  ins_add2$DebtToIncomeRatio and ins_add2$Loan_Status
S = 38937781, p-value = 0.8182
alternative hypothesis: true rho is not equal to 0
sample estimates:
         rho
-0.009295845
```

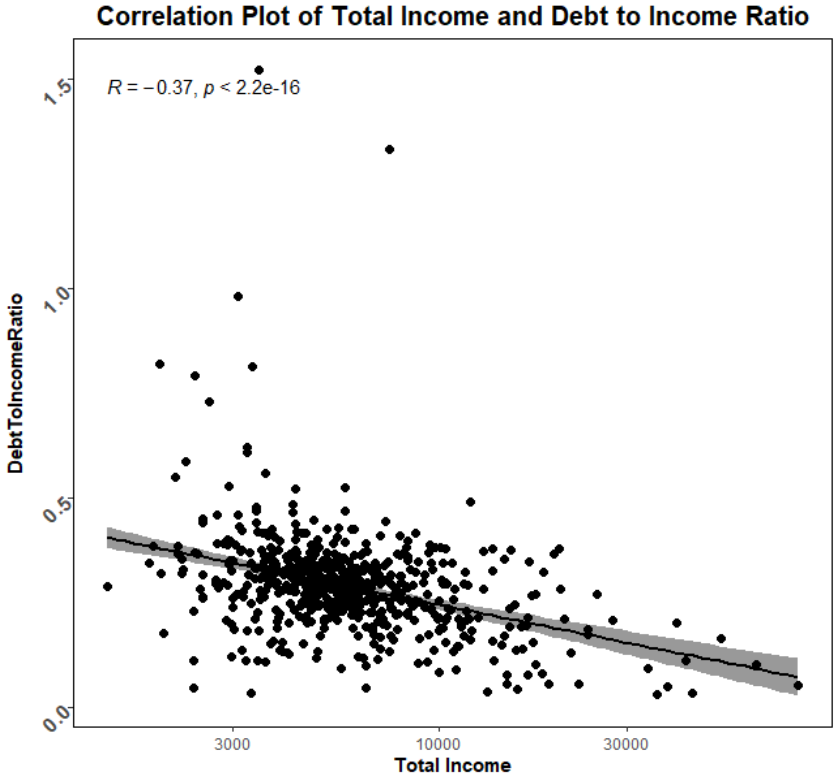
**Test of covariance and correlation between Total Income and Debt to Income Ratio:**

```
#Covariance between Total Income and Debt to Income Ratio
cov(hloan$TotalIncome, hloan$DebtToIncomeRatio)
```

**Result:**

cov = -251.7754. This value denotes that Total Income, and Debt to Income Ratio are inversely related.

rho = -0.37. This denotes negative linear relationship between Total Income and Debt to Income Ratio.




**g. Results**

The following variables were found to have significant relationship at 95% confidence level, with the loan eligibility of the applicants:

1. Credit history,
    - p-value = 2.179098e-44, cov = 0.0951, rho=0.57
2. Property area,
    - p-value = 0.002136, cov = 0.0117, rho=0.02980263
3. Civil status,
    - p-value = 0.02731, cov = 0.0197, rho =0.089
4. Educational Background,
    - p-value = 0.0333, cov = -0.0165, rho = -0.086

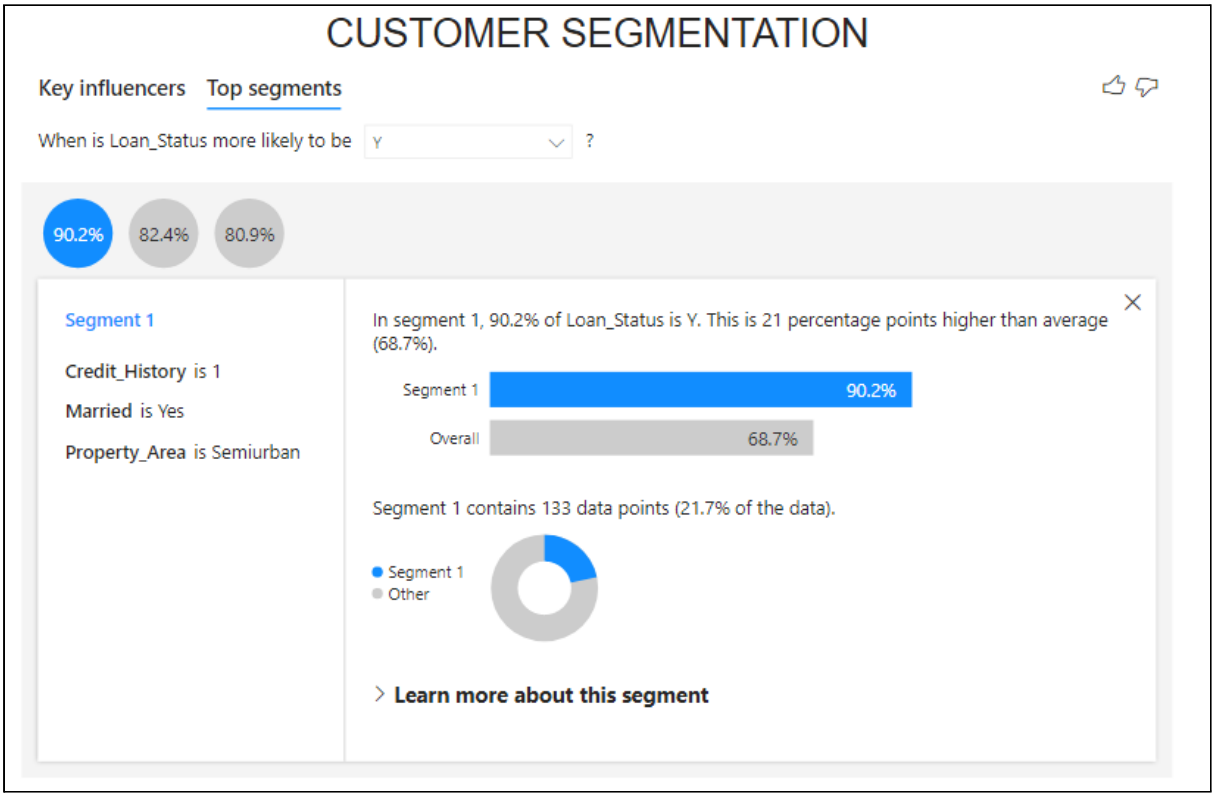
## G. Interpretations and Discussions

The chi-square residuals provided us with a closer look within the relationships and dependencies between the variables.

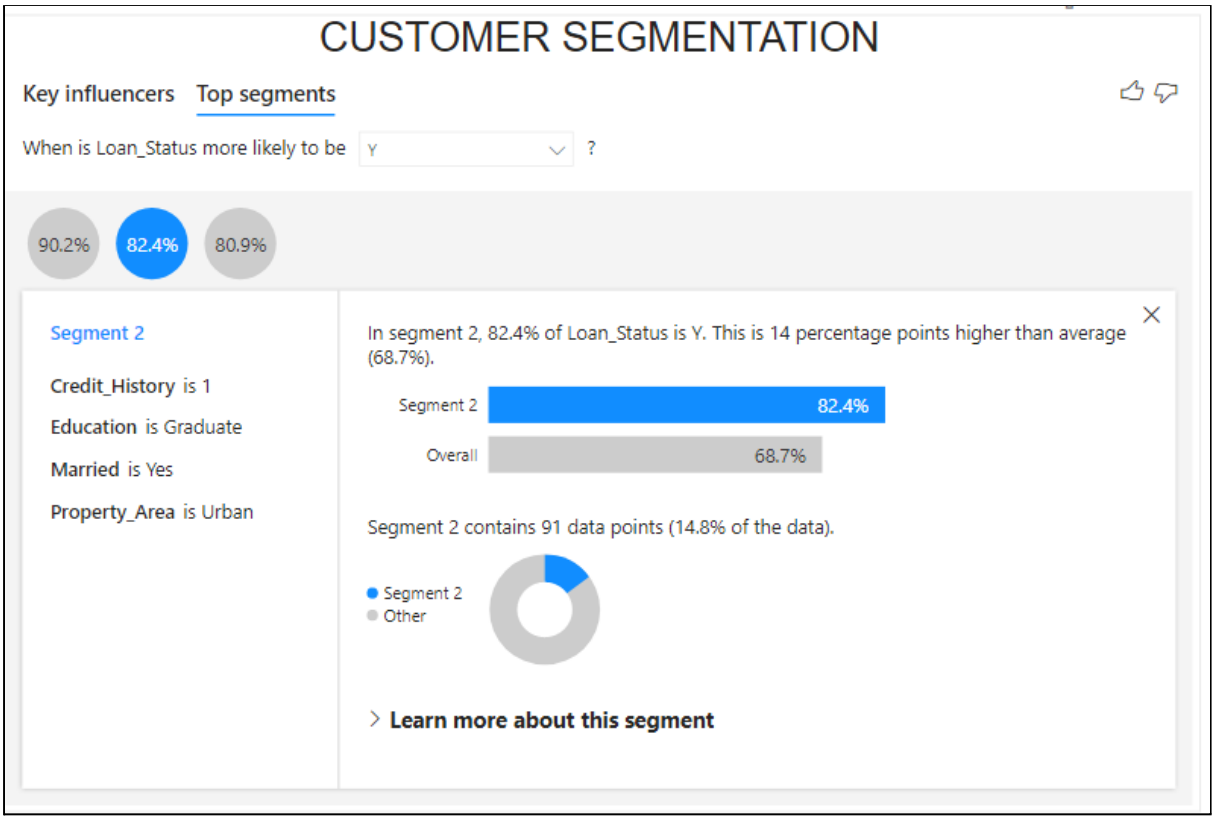
1. Credit History
    - Applicants who passed the guidelines for credit history are more likely to get loan approval.
2. Property Area
    - House loan applications in Semi-urban property areas are more likely to get loan approval.
3. Civil Status
    - Married applicants are more likely to get loan approval.
4. Educational Background
    - Applicants who graduated are more likely to get loan approval.

**Customer Segments Generated: 3**

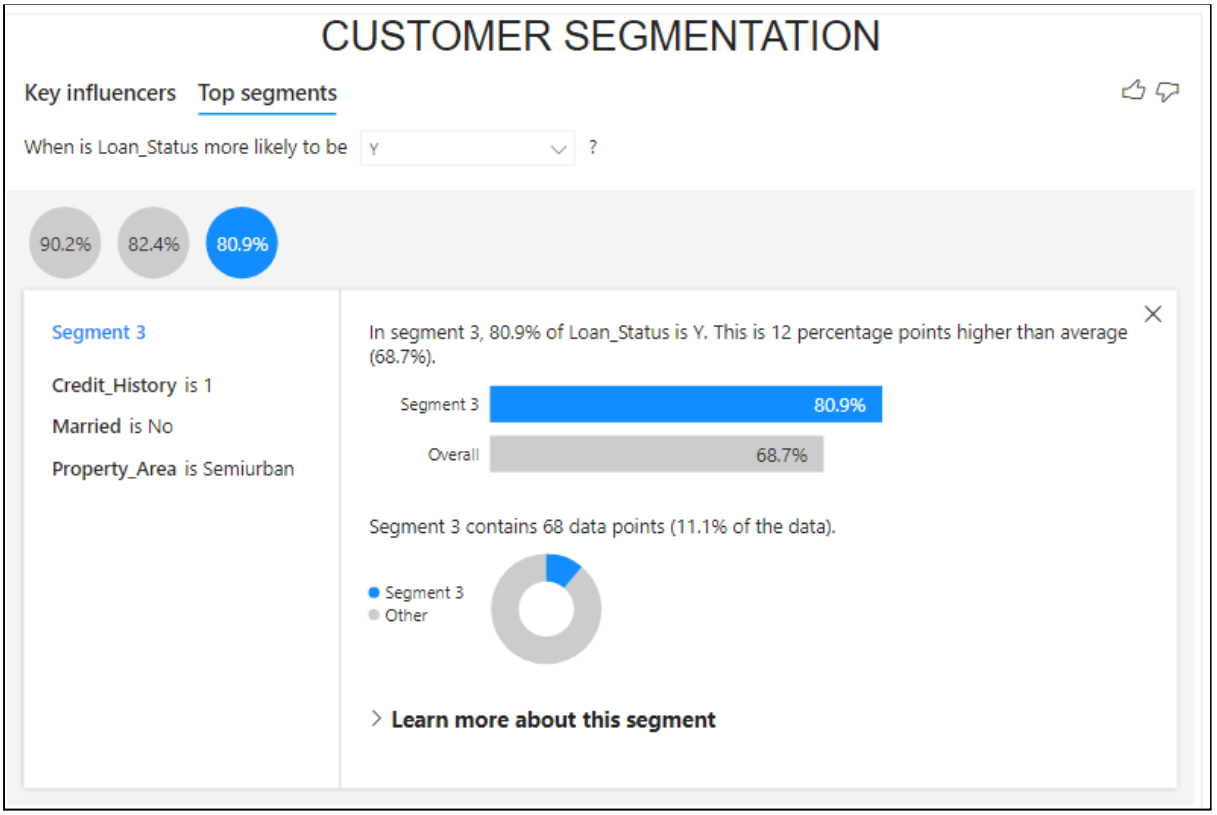
1. **Credit_History is 1, Married is Yes, Property_Area is Semiurban**



2. **Credit_History is 1, Education is Graduate, Married is Yes, Property_Area is Urban**




3. **Credit_History is 1, Married is No, Property_Area is Semiurban**



Out of all the factors affecting loan status, credit history has the strongest influence on the loan eligibility of applicants.

It is unexpected to conclude based on p-values, that capability to pay off loan in terms of (applicant income, co-applicant income, self-employment, etc.) has no significant relationship to loan eligibility.

It is counterintuitive to assume that this conclusion is true since based on common business sense, an applicant who has high income or an applicant with stable employment (not self-employed) should be more eligible for a loan.

Also, no relationship is found between loan amount and loan status where we would expect a lower loan amount to be more likely to get loan approval.